科研机构名称归一化实现*

贾君枝[1] 曾建勋[2] 李捷佳[1] 付晓梅[1]

[1] 山西大学经济与管理学院 太原 030006 [2] 中国科技信息研究所信息资源中心 北京 100038

摘要：［目的/意义］机构名称的数目多且较为繁杂，机构名称归一化可将同一机构的规范名称以及不同时段、不同表达形式的非规范名称汇集在一起，提高查询检索的查全率和查准率；有利于建立与其他系统之间的互操作，实现资源的共享。［方法/过程］在分析机构名称字符串的特点和基于K-means算法的基础上，利用编辑距离算法实现一级机构名称的初步聚类，然后利用初步聚类结果并基于TF-IDF算法计算机构名称各词项的权值，从而基于K-means算法将机构名称围绕聚类中心抱团聚簇，并对每一个簇的机构名称赋予唯一标识符。［结果/结论］该方法可实现同一机构实体不同形式的规范名称的归一，提高机构名称聚类的准确率，但对K取值、距离测度方法的选取仍有待优化。

关键词：科研机构名称 聚类 K-means

分类号：G254

DOI：10.13266/j.issn.0252-3116.2018.13.013

1 引言

科研机构是从事科学研究的高等院校、科研院所等实体对象，由于机构名称形式多样，机构合并、更替及更名频繁，机构与机构之间的关系不明确，为机构名称的识别带来了困难；此外，不同科研作者在发文时，在不同时间段对同一单位机构采用的表达方式也不同，即包括机构名称的规范名称、曾用名、简称以及错误著录形式等。从而影响到基于机构名称的信息检索、统计分析、计量评价效果。机构名称归一化旨在将同一机构实体名称的不同表达形式集中起来，建立规范名称与变异名称之间的对应关系，通过赋予机构唯一标识符的方式达到机构识别的目的。实现机构名称归一化，在数据库中检索或统计分析科研机构的学术成果时，可将同一机构的规范名称以及不同时段、不同表达形式的非规范名称汇集在一起，提高查询检索的查全率和查准率；此外，对机构规范名称赋予唯一标识符，有利于建立与其他系统之间的互操作，实现资源的共享。

对于机构名称文档的规范控制，目前国外名称规范控制工作已较为成熟，通过建立机构名称规范档、机构名称规范库等实现对机构名称别名以及各类语言名称的管理。著名的虚拟国际规范文档（Virtual International Authority File，VIAF），是2003年在柏林召开的国际图书馆协会联合会（IFLA）会议中由美国国会图书馆（LC）、德国国家图书馆和联机计算机图书馆中心（OCLC）共同发起的，其目标是通过将各个图书馆和机构的规范文档联系起来从而为同一个人或组织链接不同形式的名称[1]；规范文档链接与探索（Linking and Exploring Authority Files，LEAF），是2001年由欧洲执委会带领各国发起的，其目标是开发分布式检索系统模型架构，实现各个不同名称形式的分散记录与其规范记录的链接[2]。相比国外，国内的名称规范控制工作起步较晚，20世纪90年代我国国家图书馆开始着手中文名称规范控制。2009年，国家图书馆（NLC）、香港地区大学图书馆协作咨询委员会（JULAC-HKCAN）、台湾汉学研究中心（CSS）以及中国高等教育文献保障系统管理中心（CALIS）联合建立了CNAJDSS（China Name Authority Joint Database Search System，中文名称规范联合数据库检索系统）[3]，该系统集成了各成员单

* 本文系国家社会科学基金项目“机构规范文档结构及构建方式研究”（项目编号：15BTQ015）和国家社会科学基金重点项目“基于关联数据的中文名称规范档语义描述及数据聚合研究”（项目编号：15ATQ004）研究成果之一。

作者简介：贾君枝（ORCID：0000-0003-1486-673X），教授，博士，E-mail：junzhij@163.com；曾建勋（ORCID：0000-0002-0432-9618），主任，研究馆员，博士生导师；李捷佳（ORCID：0000-0002-2357-6315），硕士研究生；付晓梅（ORCID：0000-0002-9831-0204），硕士研究生。

收稿日期：2017-12-08 修回日期：2018-03-29 本文起止页码：103-110 本文责任编辑：刘远颖

位名称规范数据,在一定程度上实现了数据资源的共建共享;此外,中国科学院建立了其所属机构的机构名称规范库,其目标是全面构建中国科学院机构名称规范化描述,快速实现中国科学院机构规范名称和别名的登记、机构关联关系和机构名称变更关系的梳理,实现机构名称规范服务[4]。但实际上不同数据库之间存在差异,存在同一机构但采用名称不一致,缺少自动识别其更名、别名关系的机制,随着机构名称数量的增加,人工进行处理将愈为复杂,通过识别并对相应机构名称聚类是辅助人工完成机构名称规范的解决办法之一。

目前对机构名称识别的研究较多,从方法上可将其分为两类,一类是基于规则的方法[5-6],一类是基于统计的方法[7-10]。基于规则的方法主要是利用特征词触发的形式进行识别,包括从机构名称的语法性质、语义特性、组织规律及特点等进行分析并总结出相应的规则、模式和特征,通过其关键词进行识别;基于统计的方法主要是对大规模语料库进行训练,对语料库中的名称进行分析从而构建统计模型,包括组块分析技术、决策树方法、条件随机场模型、支持向量机以及隐马尔可夫模型等。对于名称归一化研究,Y. Jiang 等[11]采用一种基于归一化的压缩距离(Normalized Compression Distance)的方法实现对同一机构不同名称的聚类;杨奕虹等[12]通过编制机构多层级词表,并分析其在文献计量和科研绩效管理中的应用,解决了海量数据中机构名称归一化问题;孙海霞等[13]采用基于 K-means 算法,并借鉴基于频繁词集的文本聚类中心确定方法对机构名称进行了归一化处理,但应用 K-means 算法时对聚类中心的选择仍需改进;贤信[14]通过人工收集以及频次统计确定样本数据,采用基于 K-近邻算法与编辑距离相似度算法相结合的方法实现机构名称的归一,但其对机构名称更名关系涉及较少。综上所述,已有的机构名称归一基本实现了对机构名称别名和规范名称的聚类,但主要依靠频次统计获取规则,未将机构名称识别和归一有机地结合起来;在聚类过程中,确定聚类中心后仅采用单一方法进行聚类,且聚类过程中仅考虑了别名关系,对更名关系考虑较少。

本文在上述分析基础上,通过对词项的分析构建科研机构特征词表,发现机构名称的词性特征及组合序列,并利用构建的特征词表识别划分机构名称中的一级机构和二级机构名称;然后不进行直接聚类,将编辑距离算法、TF-IDF 方法和 K-means 算法分两部分结合应用于科研机构名称的聚类归一,从而优化了聚类效果;对于更名关系也利用作者关系做了进一步研究。

2 构建机构特征词表

特征词表构建是为了有效识别一级机构、二级机构名称,通常由表示机构性质或类型的特征词构成。

通过对科研机构名称的构成分析可以发现,科研机构名称一般以 A + B 的形式表达,并且是以 B 部分为中心语的定名型短语。A 部分一般由动词、方位词、序数词、形容词等构成且长度不定。B 部分相对较为固定,数量也较少,集中在“大学”“学院”“研究院”等,因而可以通过构建相关的特征词表识别 B 部分。而 A 部分可以通过词性标注及分析,总结出机构名称前部分的可能组合序列,在进行机构名称识别时根据序列组合规则进行匹配。

2.1 数据来源

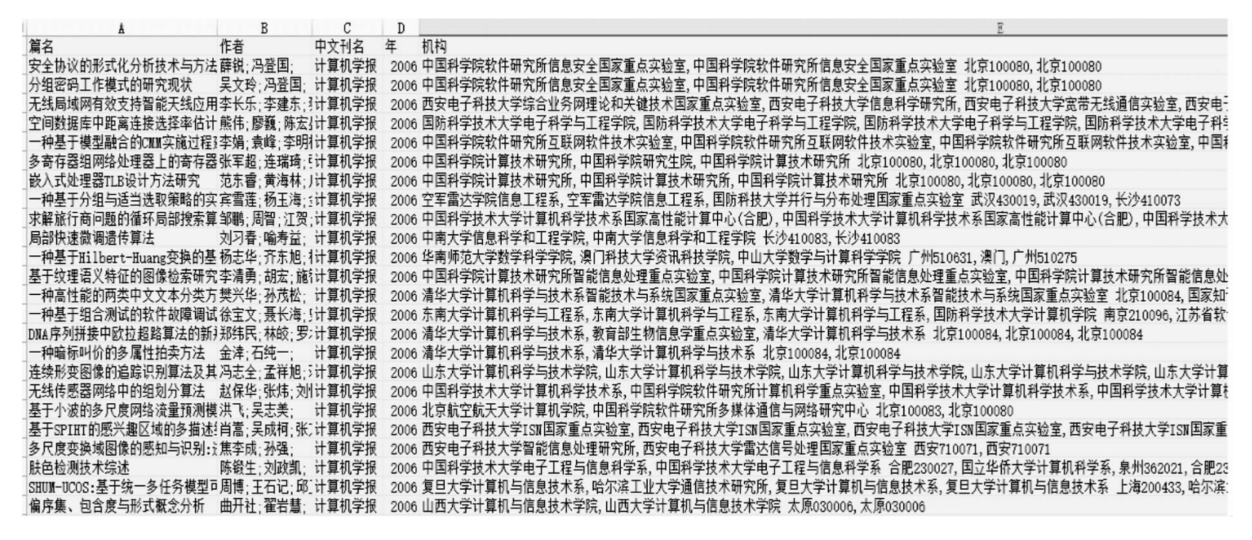
机构名称归一需要展现机构名称演化变更规律。本研究利用中国知网数据库收集样本数据,而在其中通过机构名称直接检索文献,会从不同角度影响查全率和查准率,一种是精确查找,忽略了同一机构实体其他名称的文献;另一种是模糊查找,使检索结果过于宽泛。因此在收集机构名称样本数据时,所选数据不仅需要包含机构的简称、别称等机构名称非规范形式,而且需要反映机构名称随时间的演化历程,即对时间跨度有所要求。

考虑到上述因素,本研究以学术期刊为载体,选取《图书情报工作》《计算机学报》《机械工程学报》3 种期刊的数据,收集在 2006 - 2016 年 11 间所刊论文的作者所在单位数据,涉及作者单位、作者、期刊名称及论文发表年限,如图 1 所示,从中抽取“作者单位”数据来探索机构名称的演变规律。检索出相关文献 13 839 条,而其中部分文献作者并非单一,即作者单位也存在多条数据,通过在 Excel 表中对作者单位数据进行拆分分列,保证一条记录中只包含一条单位名称数据;并且删除重复值,消除无效的噪音数据,例如“AL10”“9AB”“VA”等。经过上述对数据的预处理后,剩余有效数据 6 503 条。

2.2 机构名称的词性特征分析

词性标注分析是对自然语言处理的预处理操作,通过此可以发掘机构名称词与其相连成分的组合形式。采用 NLPIR 汉语分词系统进行分词处理及词性标注,将 6 503 条机构名称数据划分为 41 439 个词,其中主要是名词、动词、形容词,其分布情况见图 2。在分词系统词性标记集中, n 表示名词、ns 表示地名、m

贾君枝，曾建勋，李捷佳，等. 科研机构名称归一化实现[J]. 图书情报工作,2018,62(13):103 – 110.



篇名	作者	中文刊名	年	机构
安全协议的形式化分析技术与方法	薛锐;冯登国;	计算机学报	2006	中国科学院软件研究所信息安全国家重点实验室,中国科学院软件研究所信息安全国家重点实验室 北京100080,北京100080
分组密码工作模式的研究现状	吴文玲;冯登国;	计算机学报	2006	中国科学院软件研究所信息安全国家重点实验室,中国科学院软件研究所信息安全国家重点实验室 北京100080,北京100080
无线局域网有效支持智能天线应用	李长乐;李建东;	计算机学报	2006	西安电子科技大学综合业务网理论和关键技术国家重点实验室,西安电子科技大学信息科学研究所,西安电子科技大学宽带无线通信实验室,西安电
空间数据库中距离连接选择率估计	熊伟;廖巍;陈宏	计算机学报	2006	国防科学技术大学电子科学与工程学院,国防科学技术大学电子科学与工程学院,国防科学技术大学电子科学与工程学院,国防科学技术大学电子科
一种基于模型融合的CMM实施过程	李娟;袁峰;李明	计算机学报	2006	中国科学院软件研究所互联网软件技术实验室,中国科学院软件研究所互联网软件技术实验室,中国科学院软件研究所互联网软件技术实验室,中国科
多寄存器组网络处理器上的寄存器	张军超;连瑞琦;	计算机学报	2006	中国科学院计算技术研究所,中国科学院研究生院,中国科学院计算技术研究所 北京100080,北京100080,北京100080
嵌入式处理器TLB设计方法研究	范东睿;黄海林;	计算机学报	2006	中国科学院计算技术研究所,中国科学院计算技术研究所,中国科学院计算技术研究所 北京100080,北京100080,北京100080
一种基于分组与适当选取策略的实	宾雪莲;杨玉海;	计算机学报	2006	空军雷达学院信息工程系,空军雷达学院信息工程系,国防科技大学并行与分布处理国家重点实验室 武汉430019,武汉430019,长沙410073
求解旅行商问题的循环局部搜索算	邹鹏;周智;江贺;	计算机学报	2006	中国科学技术大学计算机科学技术系国家高性能计算中心(合肥),中国科学技术大学计算机科学技术系国家高性能计算中心(合肥),中国科学技术大
局部快速微调遗传算法	刘习春;喻寿益;	计算机学报	2006	中南大学信息科学和工程学院,中南大学信息科学和工程学院 长沙410083,长沙410083
一种基于Hilbert-Huang变换的基	杨志华;齐东旭;	计算机学报	2006	华南师范大学数学科学学院,澳门科技大学资讯科技学院,中山大学数学与计算科学学院 广州510631,澳门,广州510275
基于纹理语义特征的图像检索研究	李清勇;胡宏;施	计算机学报	2006	中国科学院计算技术研究所智能信息处理重点实验室,中国科学院计算技术研究所智能信息处理重点实验室,中国科学院计算技术研究所智能信息处
一种高性能的两类中文文本分类方	樊兴华;孙茂松;	计算机学报	2006	清华大学计算机科学与技术系智能技术与系统国家重点实验室,清华大学计算机科学与技术系智能技术与系统国家重点实验室 北京100084,国家知
一种基于组合测试的软件故障调试	徐宝文;聂长海;	计算机学报	2006	东南大学计算机科学与工程系,东南大学计算机科学与工程系,东南大学计算机科学与工程系,国防科学技术大学计算机学院 南京210096,江苏省软
DNA序列拼接中欧拉超路算法的新	郑纬民;林皎;罗	计算机学报	2006	清华大学计算机科学与技术系,教育部生物信息学重点实验室,清华大学计算机科学与技术系 北京100084,北京100084,北京100084
一种暗标叫价的多属性拍卖方法	金涛;石纯一;	计算机学报	2006	清华大学计算机科学与技术系,清华大学计算机科学与技术系 北京100084,北京100084
连续形变图像的追踪识别算法及其	冯志全;孟祥旭;	计算机学报	2006	山东大学计算机科学与技术学院,山东大学计算机科学与技术学院,山东大学计算机科学与技术学院,山东大学计算
无线传感器网络中的组划分算法	赵保华;张炜;刘	计算机学报	2006	中国科学技术大学计算机科学技术系,中国科学院软件研究所计算机科学重点实验室,中国科学技术大学计算机科学技术系,中国科学技术大学计算
基于小波的多尺度网络流量预测模	洪飞;吴志美;	计算机学报	2006	北京航空航天大学计算机学院,中国科学院软件研究所多媒体通信与网络研究中心 北京100083,北京100080
基于SPIHT的感兴趣区域的多描述	肖嵩;吴成柯;张	计算机学报	2006	西安电子科技大学ISN国家重点实验室,西安电子科技大学ISN国家重点实验室,西安电子科技大学ISN国家重点实验室,西安电子科技大学ISN国家重
多尺度变换域图像的感知与识别:	焦李成;孙强;	计算机学报	2006	西安电子科技大学智能信息处理研究所,西安电子科技大学雷达信号处理国家重点实验室 西安710071,西安710071
肤色检测技术综述	陈锻生;刘政凯;	计算机学报	2006	中国科学技术大学电子工程与信息科学系,中国科学技术大学电子工程与信息科学系 合肥230027,国立华侨大学计算机科学系,泉州362021,合肥23
SHUM-UCOS:基于统一多任务模型	周博;王石记;邱	计算机学报	2006	复旦大学计算机与信息技术系,哈尔滨工业大学通信技术研究所,复旦大学计算机与信息技术系,复旦大学计算机与信息技术系 上海200433,哈尔滨
偏序集、包含度与形式概念分析	曲开社;翟岩慧;	计算机学报	2006	山西大学计算机与信息技术学院,山西大学计算机与信息技术学院 太原030006,太原030006

图 1　预处理前数据

表示数词、vn 表示名动词、cc 表示并列连词、b 表示区别词。从统计结果可以看出,机构名称中名词占绝大多数,对机构名称的组合序列进行总结,主要分为 4 种类型:①名词 + 名词。比如"中国/ns 人民/n 大学/n 新闻/n 学院/n",此构成类型占了绝大多数。②动词 + 名词。比如"武汉/ns 大学/n 新闻/n 与/cc 传播/vn 学院/n"。③形容词 + 名词。比如"山东/ns 大学/n 公共/b 卫生/an 学院/n"。④序数词 + 名词。比如"广东省/ns 岭南/ns 工商/n 第一/m 技师/n 学院/n"。

通过词性分析,在下一步构建特征词表时,主要对名词进行判断,可以根据序列组合形式辅助判断是否为特征词汇。

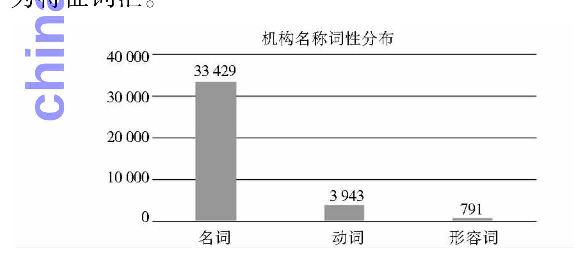


图 2　机构名称词性特征

2.3　确定特征词表

除了对词性分布进行统计,对词频也进行了统计。共划分为 82 946 个词汇,频次最高的为"大学",共出现 3 411 次;其次为"学院",共出现 2 727 次。说明"大学"和"学院"两个词很大程度上可能是科研机构的特征词。"大学"一般都是作为高校一级机构的特征词;"学院"的频次也较高,在科研机构中,"学院"既可以作为一级机构存在,也可以作为二级机构存在,如"太原师范学院地理科学学院"。此外,"工程""信息""科学"等词的词频也较高,通过对数据的调研,此类型大多为表示行业方向、专业类型的词,与原始数据类型息息相关;收集的原始数据不同,所统计的词频也随之发生变化。

构建机构名称特征词表,需要选取频率较高的词,即选用泛词,因而采用高频低频词界定公式计算选出高频词汇。高频低频词阈值采用 J. C. Donohue 在文献[15]中利用齐普夫定律推导出的高频低频词界分公式,如式(1)所示:

$$T = (-1 + \sqrt{1 + 8 \times I_1})/2 \qquad 式(1)$$

I_1 为词频为 1 的关键词的个数,T 指高频词中的最低频次值,即高频、低频词的词频临界值。本统计中,I_1 = 1 090,计算得 T = 46,最终确定本研究的高频低频词阈值为 46 次。选取出现频次大于 46 次的词作为高频词,并对这些高频词进行人工审核,将符合条件的词汇作为科研机构的特征词存储至特征词库中,最终确定了 11 个特征词,如表 1 所示:

表 1　机构名称特征词表

特征词	词频
大学	3 411
学院	2 727
图书馆	945
实验室	754
系	714
研究所	659
中心	485
科学院	269
研究院	248
公司	112
集团	82

3 机构名称层级识别

在进行机构名称归一前，需要识别其一级机构和二级机构。借助分词系统只是对于机构名称进行了分词和词性标注，无法区分一级机构和二级机构。

机构名称的识别包括一级机构名称、二级机构名称等各级机构名称的识别，根据已构建的机构特征词表，从左向右正向遍历每个机构名称字符串，然后为一级机构、二级机构等设置符号，并按照先后顺序在与特征词表匹配成功时标注相应的符号将其区分开，即实现一、二级机构名称的识别[14]。

处理流程如下：对机构数据进行预处理后，首先将机构数据与特征词表进行精确匹配，对每一次匹配成功的标注设定的符号。如"山西大学经济与管理学院"，则标注为"山西大学#经济与管理学院#"，即将"山西大学"视为一级机构，"经济与管理学院"视为二级机构。对于包含特征词较多的，比如"北京师范大学管理学院信息管理与信息系"，将其标注为"北京师范大学#管理学院#信息管理与信息系#"，即将"北京师范大学"视为一级机构，"管理学院"视为二级机构，"信息管理与信息系"视为三级机构。对于一些特殊机构名称，如"山西大学商务学院"，标注为"山西大学#商务学院#"，事实上其本身为一级机构，标注后的划分不符合实际情况，对于这一类问题，将在第四部分设计解决。通过上述流程，识别结果如表 2 所示：

表 2 机构名称层次识别结果（示例）

一级机构	一级机构绑定	二级机构	三级机构
XX 科学院	研究所、研究中心、大学、研究生院等	实验室、中心、学院等	XX 部、XX 系等
XX 大学	XX 学院、分校等	学院、系、图书馆、研究所、实验室等	XX 系、XX 部、办公室等
XX 学院	分校等	学院、系、图书馆、研究所、实验室等	XX 系、XX 部、办公室等

4 机构名称归一

机构名称归一化旨在将同一机构实体名称的不同表达形式集中起来，建立规范名称与变异名称之间的对应关系，通过赋予机构唯一标识符的方式达到机构识别的目的。

基于计算字符串相似度的算法对数据中一级机构名称进行初步聚类，将同一个一级机构归入同一数据团。按照计算所依据的特征字符串相似度计算方法可以划分为：基于字面相似的方法、基于统计关联的方法、基于语义相似的方法以及综合字面、语义和统计关联特征的多层特征方法[16]。其中编辑距离算法可以根据设定相似度阈值对字符串进行分组，并且应用广泛，发展较为成熟，可以满足机构名称初步聚类的要求。再基于 TF-IDF 算法并通过借鉴 K-Means 聚类算法的中心思想将机构名称围绕聚类中心抱团聚簇，利用初步聚类结果进行一级机构名称下的小范围聚类，即实现同一机构实体不同形式名称的聚合。

4.1 初步聚类

编辑距离算法（levenshtein distance）是 V. I. Levenshtein 于 1966 年提出的[17]，编辑距离是指由源字符串转变为目标字符串所需要的最小编辑操作的次数，编辑操作包括"插入""删除""替换"3 种类型。

计算相似度时，先取两个字符串长度的最大值，然后通过式(2)计算相似度(Sim)：

$$Sim = 1 - (编辑次数/最大值) \quad 式(2)$$

当由源字符串转换为目标字符串时，源字符串不需要进行任何编辑操作就转换为目标字符串，即编辑距离为 0，相似度为 100%，则两个字符串完全相似；当由源字符串转换为目标字符串时，源字符串中每个字符全部转换为目标字符串中的字符方可相同，即相似度为 0%，则两个字符串之间没有相似性。因而需要设定一个介于 0% 和 100% 之间的相似度阈值，对源字符串和目标字符串进行编辑距离计算后判断相似度是否满足预先设置的阈值 Y，从而判断是否进行聚类。

将识别后的一级机构名称筛选出来作为数据样本。初步聚类算法如下：首先设定编辑距离算法的相似度阈值 Y，然后从数据样本中将第一条数据作为第一个团的聚类中心，接着将剩余数据依次与第一条数据进行相似度计算，如果相似度小于阈值 Y 不作处理；如果相似度大于等于阈值 Y，则将这条数据与第一条数据分入同一个团中。然后从相似度小于阈值 Y 的数据中再次选取第一条数据作为聚类中心，重复循环上述操作，直至所有数据完成编辑距离计算，即可将所有一级机构名称数据划分为若干个机构名称数据团，实现步骤见图 3。

4.2 基于 TF-IDF 和 K-means 算法的机构名称归一

依据编辑距离算法完成对一级机构名称的初步聚类后，在每一个一级机构名称数据团下，首先计算各个机构名称的 TF-IDF（term frequency-inverse document

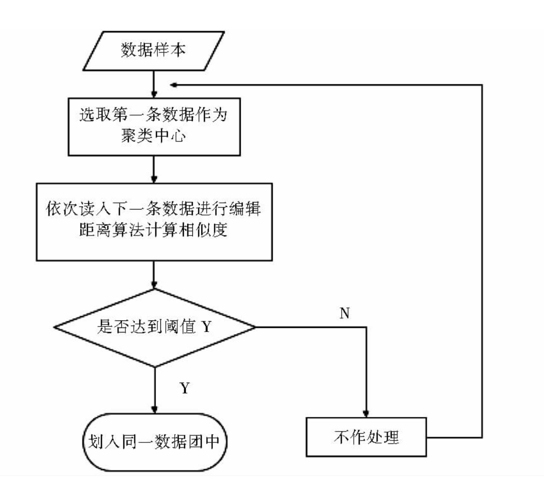


图 3　初步聚类流程

frequency)值,再采用 K-means 算法将其一、二级机构名称绑定进行迭代聚类,缩小聚类范围,实现机构名称数据的归一。

TF-IDF 是指 TF ∗ IDF,可以评估一词项在文档集合中的重要程度[18]。TF 即词频(term frequency),表示某词项在文档中出现的频率;IDF 即反文档频率(inverse document frequency) = $\log_2(N/DF)$,其中 DF(document frequency)表示包含某词项的文档数量。TF-IDF 的主要思想是:若某词项在某文档中出现的频率高,而在整个文档集合中其他文档中出现的频率低,则该词项对文档具有区别意义,可以对文档进行分类;若包含某此项的文档越少,IDF 越大,则该词项对不同类别文档的区分度越高。算法流程如下:首先对文档进行分词,并去除停用词;然后统计各个词项在单个文档中出现的次数和文件集合中词项出现的次数;最后计算得出其 TF-IDF 值。

传统的 K-means 聚类算法是预先设定聚类中心,指定类别数后对样本集合进行聚类,并且采用迭代更新的算法向目标函数值减少的方向进行,使目标函数值取得极小值,达到较优的聚类效果[19]。通过 TF-IDF 可以衡量每个词项在文档中的重要程度,将其嵌套 K-means 算法流程如下:首先从数据对象中随机选取若干个元素作为 K 个簇的初始聚类中心,将其 TF-IDF 值代入,分别计算剩下的其他数据与各个簇的聚类中心的距离,将数据赋给与其距离最近的簇;然后根据聚类结果,重新计算即调整每个簇的聚类中心,将聚类的中心移到聚类的几何中心(均值)处;反复迭代,直到聚类中心不再移动,即算法收敛。在计算距离时,本文采用 K-means 算法中常用的欧几里得距离[20],即两个元素在欧式空间中的集合距离:

$$D(X_i,Y_i)=\sqrt{((x1-y1)^2+(x2-y2)^2+\cdots+(xn-yn)^2)}$$

式(3)

其中 X,Y 分别代表文档,X_i,Y_i 是每个文档中词项的 TD-IDF 值。

聚类完毕后,分别对每一个簇的机构名称赋予唯一标识符 ID。

如图 4 所示:

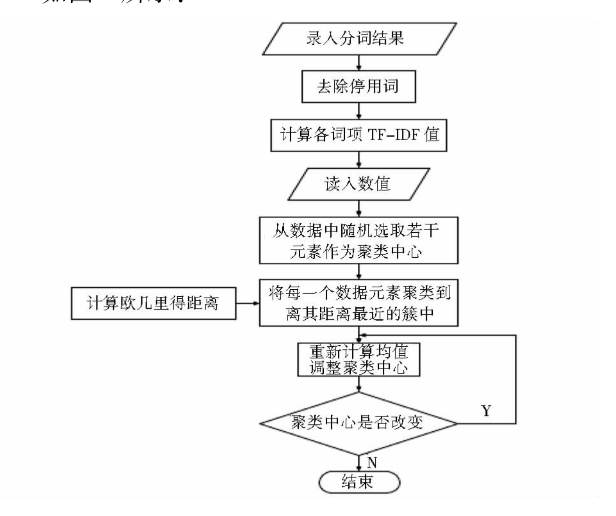


图 4　聚类流程

5　机构更名关系的提取

聚类方法适用于识别机构的别名,即名称之间存在一定的相似性。除此之外,实际机构名称数据中还存在更名情况,比如“中北大学”原名为“华北工学院”,此类更名关系以及其更名规律难以从词性规则归纳总结,使得机构名称无法从语词的相似性建立关联。因而利用机构数据信息中作者的共现率挖掘机构之间的关联关系,此处共现率是指所比较的两个机构中共同作者(作者交集)占作者总数(作者并集)的比值。由于提取机构名称的更名关系需要计算机构之间的作者交集,而计算作者交集就要考虑交集数达到多少才判定两个机构之间有关联,即首先需要提前设置共现率阈值;经过机构名称数据聚类之后,每一个机构名称簇所代表的机构都有其相应的作者集合,然后通过计算机构名称簇所对应的作者集合之间的共现作者数提取更名关系。

实现方法如下:设置作者共现率的阈值 X;从机构信息中的第一个作者集合开始,依次迭代提取每个作者集合和其他所有作者集合的交集,并计算出相应的共现率;若计算所得的共现率小于阈值 X,则不作处理;若共现率大于等于阈值 X,则判定两个机构名称之间存在更名关系,并将两个机构的唯一标识符 ID 合并

（采用时间较近的名称的 ID 赋予），对时间较远的标注曾用名。更名关系提取的原理如图 5 所示：

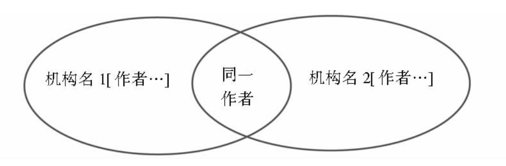


图 5 提取更名关系原理

6 实验分析

本实验首先使用 MyEclipse 开发工具，用 Java 语言编写程序完成机构名称层级识别、编辑距离初步聚类以及基于 TF-IDF 和 K-means 算法的文本聚类等操作；然后采用 C 语言实现机构名称更名关系的提取，最终实现机构名称的归一化研究。

6.1 机构名称层次识别

在本次实验中，采用预处理后的机构名称样本数据进行操作。识别结果如图 6 所示（以北京大学相关机构名称为例）：

543	北京大学#
544	北京大学#CALIS管理中心#
545	北京大学#CASHL管理中心#
546	北京大学#法制信息中心#
547	北京大学#高可信软件技术教育部重点实验室#
548	北京大学#高可信度软件技术教育部重点实验室#
549	北京大学#光华管理学院#
550	北京大学#机器感知与智能教育部重点实验室#
551	北京大学#计算语言学教育部重点实验室#
552	北京大学#计算语言学研究所#
553	北京大学#计算机科学与技术系#
554	北京大学#计算机科学技术系#
555	北京大学#计算机科学技术研究所#
556	北京大学#计算中心#

图 6 机构名称识别结果（部分）

6.2 机构名称归一

6.2.1 基于编辑距离算法的初步聚类 根据识别结果，分离出一级机构并进行预处理，即排序去重。去重后得到 2 465 条一级机构名称数据。运用编辑距离算法进行初步聚类，其中经过对数据的分析，相似度阈值设置为 85% 较为准确。初步聚类结果如图 7 所示（以河北地区为例）。

716	一级机构名称	一级机构名称	编辑距离	相似度
717	河北北方学院	河北北方学院:	ld: 0	sim: 1.0
718	河北大学	河北大学:	ld: 0	sim: 1.0
719	河北钢铁集团	河北钢铁集团:	ld: 0	sim: 1.0
720	河北工程大学	河北工程大学:	ld: 0	sim: 1.0
721	河北工业大学	河北工业大学:	ld: 0	sim: 1.0
722	河北工业职业技术学院	河北工业职业技术学院:	ld: 0	sim: 1.0
723	河北工业职业技术学院	河南工业职业技术学院:	ld: 1	sim: 0.9
724	河北建材职业技术学院	河北建材职业技术学院:	ld: 0	sim: 1.0
725	河北金融学院	河北金融学院:	ld: 0	sim: 1.0
726	河北经贸大学	河北经贸大学:	ld: 0	sim: 1.0
727	河北科技大学	河北科技大学:	ld: 0	sim: 1.0
728	河北科技师范学院	河北科技师范学院:	ld: 0	sim: 1.0
729	河北理工大学	河北理工大学:	ld: 0	sim: 1.0
730	河北联合大学	河北联合大学:	ld: 0	sim: 1.0
731	河北旅游职业学院	河北旅游职业学院:	ld: 0	sim: 1.0
732	河北民族师范学院	河北民族师范学院:	ld: 0	sim: 1.0
733	河北农业大学	河北农业大学:	ld: 0	sim: 1.0
734	河北软件职业技术学院	河北软件职业技术学院:	ld: 0	sim: 1.0
735	河北省电力公司	河北省电力公司:	ld: 0	sim: 1.0
736	河北省科学技术情报研究院	河北省科学技术情报研究院:	ld: 0	sim: 1.0

图 7 一级机构名称初步聚类结果（部分）

从图 7 中可以看出，一级机构之间相似度达到 85% 的已筛选出来。标红部分为阈值达到 85% 的特殊情况，主要有以下几类：

（1）两者指同一机构，但因其命名等原因著录失误，经过初步聚类后实现抱团。例如：地名问题："广东外语外贸大学"和"广州外语外贸大学"，实际均为"广东外语外贸大学"；介词问题："国防科学与技术大学"和"国防科学技术大学"，实际均为"国防科学技术大学"；省市标识问题："山西旅游职业学院"和"山西省旅游职业学院"实际均为"山西旅游职业学院"，"上海交通大学"和"上海市交通大学"实际均为"上海交通大学"。

（2）两者并非同一机构，但因机构名称之间编辑距离较少，被视为相似。此类情况在后期一、二级机构名称绑定归一化重新聚类。①地名前缀相似："北京航空航天大学"和"南京航空航天大学"、"河南财经政法大学"和"中南财经政法大学"等；②修饰词相似："哈尔滨工业大学"、"哈尔滨工程大学"和"哈尔滨商业大学 "；③同一附属机构统一编名："中国航天科工集团第六研究院"和"中国航天科工集团第三研究院"。

对于这些特殊情况，进行抱团聚簇时，若两个机构名称相似且机构名称数目相当，进行 TF-IDF 计算后区分度较高，通过 K-means 聚类可将其划分为相应的两个机构簇。但当两个机构名称相似却机构名称数目差异较大时，并不能满足划分的需求，需考虑利用作者交集信息进行分析。提取两个相似机构名称的作者信息，依据计算更名关系时采用的共现率判断是否为同一机构。若达到设置的阈值，即将其视为同一机构进行相应的聚类；若低于阈值，则依据人工判断进行划分，再进行聚类。

6.2.2 基于 TF-IDF 和 K-means 算法的聚类 根据一级机构数据团的聚类结果，对每一个一级机构下的二级机构名称首先计算其分词后各词项的 TF-IDF 值，然后基于 K-means 算法，将一、二级机构名称绑定进行抱团聚簇。

图 8 和表 3 是对一级机构"山西大学"下的机构名称进行聚类分析的结果。

根据聚类结果，提取每个机构名称相关文献的出版日期并进行排序，可建立其别名关系。将日期最新的作为中心名称，其他机构名称作为别名，建立机构名称别名映射表。如 "0300101" 即将"山西大学经济与

贾君枝，曾建勋，李捷佳，等.科研机构名称归一化实现[J].图书情报工作,2018,62(13):103－110.

```
-----------------
山西 大学 管理 学院
山西 大学 计算机 与 信息 技术 学院
山西 大学 经济 与 工商 管理 学院
山西 大学 经济 与 管理 学院
山西 大学 信息 管理 系
山西 大学 信息 管理 系
山西 大学 经济 与 管理 学院

-----------------
山西 大学 人事处
-----------------
山西 大学 商务 学院
山西 大学 商务 学院 图书馆
山西 大学 商务 学院 信息 学院
-----------------
```

图8　机构名称归一结果示例

表3　赋予机构名称唯一标识符

ID	机构名称
300101	山西大学管理学院
	山西大学经济与工商管理学院
	山西大学经济与管理学院
	山西大学信息管理系
	山西大学计算机与信息技术学院
300102	山西大学人事处
300103	山西大学商务学院
	山西大学商务学院图书馆
	山西大学商务学院信息学院

管理学院”作为该簇的中心名称，将“山西大学管理学院”“山西大学经济与工商管理学院”“山西大学信息管理系”“山西大学计算机与信息技术学院”存储至其别名表中建立映射。

在聚类结果中，也存在聚类错误的情况，如“山西大学计算机与信息技术学院”与“山西大学经济与管理学院”并不是同一个机构，采用 TF-IDF 计算，“计算机与信息技术”和“信息管理系”中“信息”一词的频次较高，影响了聚类效果，因而准确率并不能达到100%。

6.3　机构更名关系的提取

整合作者信息数据后，通过对数据的考察将共现率阈值设置为15%。应用 C 语言代码测试，结果见图9，提取“华北工学院”和“中北大学”对应的作者集合，并计算其共现率。两个机构作者集合的共现率达到了15%，因而认为两个机构之间存在更名关系，由于“中北大学”相关文献的出版日期新于“华北工学院”，将两个机构名称合并入同一个簇中，并将“中北大学”的唯一标识符 ID 赋予该簇。

6.4　实验结果评析

根据具体的实践应用需求，观察聚类的结果是否与预期结果相符是检验一个聚类是否有效的方法。本研究目标是实现同一机构不同表达形式的自动归一，

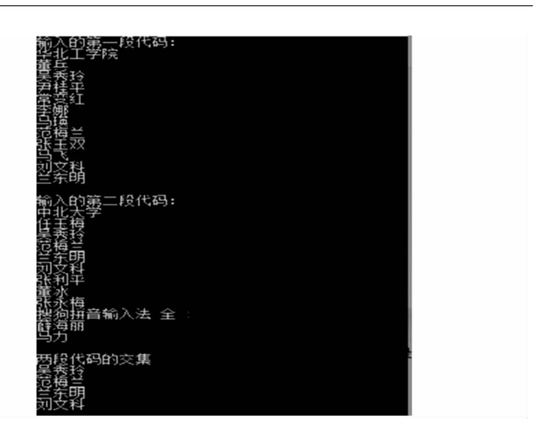


图9　机构更名关系提取

评价指标主要包括聚类效率和聚类效果，聚类效率涉及到算法的采用、应用程序、机器设备等，聚类效果主要考虑准确率(precision)和召回率(recall)两个指标，定义如下：

$$准确率：P = A/(A+B)$$

$$召回率：R = A/(A+C)$$

其中，A 表示在实验机构数据中正确聚类的机构名称数；B 表示在实验机构数据中错误聚类的机构名称数；C 表示在实验机构数据中未正确聚类，但确实为该机构类别的机构名称数。将本文研究方法的聚类结果进行人工验证，在一级机构名称归一的基础上，选取部分一级机构名称，对其下的二级机构进行聚类，准确率达到了80%以上，召回率达到了75%以上，聚类效果较好，但仍存在一些问题，有待进一步对算法进行优化。

7　结语

随着科学技术的不断发展，科研人才不断涌出。而其所在机构也由于著录缺失、演化更替、岗位变动等产生不同的表达形式，在对科研实体进行统计评价时较为复杂，需要借助一定的方式构建机构规范文档，集中其所有的表达形式。本文改变了以往直接采用 K-means 算法进行聚类的算法，在聚类前的预处理阶段，依据构建的特征词表进行机构名称识别，划分一级机构与二级机构等层级。通过应用编辑距离算法对一级机构进行初步聚类，限定了聚类范围，从而缩短了聚类时间，提高了算法运行效率。利用 TF-IDF 识别区分特征，一定程度上提高了聚类的准确率和召回率。通过这两部分聚类，优化了单独聚类的效果。将机构名称数据抱团聚簇，可以挖掘其各名称之间的关系，建立相应的规则，便于统计科研成果以进行定量研究，在实践中减少了人工构建规范库的工作量。

此外，选取机构数据时，只选取了近11年的文献

数据,处理的数据量较少,科研机构的发展需要一定的时间,而机构 11 年间更名的较少,如从“学院”转为“大学”、多个院校进行合并等,因而分析机构名称演化及进行归一,还是需要时间跨度较大的数据进行分析,本文在此处尚有欠缺,在分析机构别名和更名关系时具有一定的局限性,需要进一步考虑扩大数据量。应用 K-means 算法聚类时,K 值的确定仍对结果有影响,需要借助机器学习方法或基于统计的方法对大量数据进行训练,对取值进行优化;对于不同的机构名称数据,应当从各个角度考虑,不断调整 K 值使得结果最优。这些问题还有待进一步解决。

参考文献:

[1] VIAF[EB/OL]. [2017-03-06]. http://www.oclc.org/en/viaf.html.
[2] LEAF - linking and exploring authority files[EB/OL]. [2018-03-27]. http://www.dlib.org/dlib/september01/09inbrief.html.
[3] 中文名称规范联合数据库检索系统[EB/OL]. [2017-12-01] http://cnass.cccna.org/jsp/index.jsp.
[4] 中国科学院机构名称规范库[EB/OL]. [2017-03-08]. http://irsr.llas.ac.cn/institution/.
[5] 张小衡, 王玲玲. 中文机构名称的识别与分析[J]. 中文信息学报, 1997, 11(4):21-32.
[6] 沈嘉懿, 李芳, 徐飞玉,等. 中文组织机构名称与简称的识别[J]. 中文信息学报, 2007, 21(6):17-21.
[7] 陈霄, 刘慧, 陈玉泉. 基于支持向量机方法的中文组织机构名的识别[J]. 计算机应用研究, 2008, 25(2):362-364.
[8] 俞鸿魁, 张华平, 刘群,等. 基于层叠隐马尔可夫模型的中文命名实体识别[J]. 通信学报, 2006, 27(2):87-94.
[9] 叶琳莉, 黄日茂. 结合决策树方法的中文机构名称识别[J]. 福建电脑, 2007(12):184-184.
[10] 尹继豪, 樊孝忠, 赵攀超,等. 基于组块分析技术的中文机构名称识别[J]. 哈尔滨工程大学学报, 2006, 27(S1):466-470.
[11] JIANG Y, ZHENGH T, WANG X, et al. Affiliation disambiguation for constructing semantic digital libraries[J]. Journal of the American Society for Information Science &Technology, 2011, 62(6):1029-1041.
[12] 杨奕虹, 李雅萍, 张立丽,等. 机构多层级词表的编制及在文献计量评价与科研绩效管理中的应用[J]. 数字图书馆论坛, 2013(6):57-63.
[13] 孙海霞, 李军莲, 吴英杰. 基于 K-means 的机构归一化研究[J]. 医学信息学杂志, 2013, 34(7):41-44.
[14] 贤信. 机构规范文档结构及构建方式研究[D]. 北京:中国科学技术信息研究所, 2015.
[15] DONOHUE J C. Understanding scientific literatures: abibliometricapproach[M]. Cambridge: The MIT Press, 1973: 49-50.
[16] 章成志. 基于多层特征的字符串相似度计算模型[J]. 情报学报, 2005, 24(6):696-701.
[17] LEVENSHTEIN V I. Binary codes capable of correcting deletions, insertions and reversals[J]. Soviet physics doklady, 1966, 10(1):707-710.
[18] 吴军. 数学之美[M]. 北京:人民邮电出版社, 2014.
[19] 李飞, 薛彬, 黄亚楼. 初始中心优化的 K-Means 聚类算法[J]. 计算机科学, 2002, 29(7):94-96.
[20] 何晓群. 多元统计分析[M]. 北京:中国人民大学出版社, 2012.

作者贡献说明:

贾君枝:负责论文选题指导,研究框架设计,论文修改;
曾建勋:负责论文研究内容指导;
李捷佳:负责资料收集与整理,论文写作;
付晓梅:负责数据收集,论文写作。

Realization of Research Institution Name Normalization

Jia Junzhi[1] Zeng Jianxun[2] Li Jiejia[1] Fu Xiaomei[1]
[1] School of Economics and Management, Shanxi University, Taiyuan 030006
[2] Institute of Scientific and Technical Information of China, Beijing 100038

Abstract: [**Purpose/significance**] Institution names are numerous and complicated. The normalization of institution names brings the authoritative name and the informal ones (both at different times and in different ways of expression) of the same institution together, enhancing comprehensiveness and accuracy of searches, promoting interoperability with other systems, and thus realizing resource sharing. [**Method/process**] Based on the analysis of institution names' characteristic and K-means algorithm, this paper utilizes the edit distance similarity algorithm to achieve name normalization of institution names. Then uses TF-IDF to calculate the weight of each item, around the cluster center to normalize institution name based on K-means algorithm and gives the unique identifier to every cluster. [**Result/conclusion**] It achieves name normalization of the same institution name in different forms. And it improves the precision of institution name cluster, but the choice of K value and distance measurement method still needs to be optimized.

Keywords: research institution name　cluster　K-means